Journal of the Korean Data & Information Science Society
2015, **26**(6), 1259–1269

http://dx.doi.org/10.7465/jkdi.2015.26.6.1259
한국데이터정보과학회지

텍스트 마이닝을 활용한 영화흥행 예측 연구[† ‡]

이상훈[1] · 조장식[2] · 강창완[3] · 최승배[4]

[1](주) 온솔 커뮤니케이션 · [2]경성대학교 응용통계학과 ·
[34]동의대학교 데이터정보학과

접수 2015년 7월 21일, 수정 2015년 9월 4일, 게재확정 2015년 9월 16일

요 약

최근 빅 데이터는 학계에서 키워드로 자리매김을 하고 있다. 빅 데이터의 유용성은 학계뿐만 아니라 정부, 지자체 그리고 기업체까지 파급되고 있고, 빅 데이터 속에서 유용한 정보를 도출해 내기 위해 노력하고 있다. 본 연구에서는 영화에 대한 리뷰를 가지고 텍스트 마이닝 (text mining)을 이용한 빅 데이터 분석을 수행한다. 본 연구의 목적은 포털 사이트 'D'사와 영화진흥위원회의 영화에 대한 리뷰 데이터, 그리고 고객들의 평점평균 (score)과 스크린 수 (screen number)를 설명변수로 사용하고, 영화 흥행 여부를 종속변수로 하여 로지스틱 회귀분석을 통한 영화 흥행 예측 모형을 제안하는 것이다. 분석결과, 본 연구에서 제안한 예측모형의 정분류율은 95.74%로 얻어졌다.

주요용어: 감성분석, 정분류율, 텍스트 마이닝, 특이값 분해.

1. 서론

현대의 사회는 발달된 과학문명으로 인해 여러 가지 센서와 통신기기가 늘어나면서 방대한 데이터가 발생하고 있다. 최근 이러한 방대한 데이터 속에서 유용한 정보를 얻어내기 위한 방법으로 빅 데이터 분석이 다양한 분야에서 관심을 가지고 연구되어 지고 있다. 이는 과거와는 달리 최근의 정보화 기술이 발달하게 됨에 따라 방대한 양의 데이터를 저장할 수 있게 되었고, 이를 처리·분석하는 기술의 발달에 기인한다고 할 수 있다. 다양한 분야에서 얻어지는 빅 데이터는 정형 또는 비정형 형태로 발생되고 데이터 크기 역시 방대하다.

비정형화 데이터의 종류는 사회네트워크 서비스 (social network service; SNS), 각 종 사회미디어 (social media), 그리고 어떠한 생각이나 상태를 나타내는 이미지 등 여러 가지 형태로 존재한다. 비정형화 형태의 빅 데이터는 다양한 도구 (예를 들면, 하둡)를 이용하여 정형화하여 분석을 수행한다. 방대한 데이터들 속에서 유용한 정보를 얻기 위한 분석 방법은 데이터의 형태에 따라 여러 가지 분석기법이 있지만, 대표적인 분석 방법으로 데이터 마이닝 (data mining)이 있다. 데이터 마이닝의 종류에는 텍스트 마이닝, 웹 데이터 마이닝 (web data mining), 공간 데이터 마이닝 (spatial data mining) 등이 있다. 또한 텍스트 마이닝의 일부분인 오피니언 마이닝 (opinion mining)은 어떤 객체 (상품, 영화 등)에 대한

† 이 논문은 2015년도 동의대학교의 연구비에 의하여 수행되었음 (2015AA098).
‡ 이 논문은 공동저자 이상훈의 석사학위 논문을 재구성한 것임.
[1] (153-803) 서울시 금천구 가산동 디폴리스 지식산업센터 A동 1506호, 사원.
[2] (608-736) 부산광역시 남구 수영로 309 번지, 경성대학교 정보통계학과, 교수.
[3] (614-714) 부산광역시 부산진구 가야동 산 24번지, 동의대학교 데이터정보학과, 교수.
[4] 교신저자: (614-714) 부산광역시 부산진구 가야동 산 24번지, 동의대학교 데이터정보학과, 교수.
E-mail : csb4851@deu.ac.kr

긍정과 부정에 대한 감정에 대해 중요성을 두고 있다. 이는 감성 분석 (sentiment analysis)으로 불리기도 하며 이슈, 사건, 토픽 등 이들의 여러 속성에 대한 사람들의 의견, 태도, 감정 등을 분석한다. 텍스트 마이닝의 경우, 분석대상은 문서 또는 웹상의 문자와 같은 텍스트로써 이러한 텍스트들에 대한 의미를 파악하여 유용한 정보를 획득하는 등 다양한 연구가 수행되어 왔다.

이에 대한 대표적인 연구로는 Kim과 Oh (2009)는 온라인 고객들의 리뷰를 효과적으로 사용하기 위해 시장세분화 개념을 도입하여 텍스트를 범주화시켜 연구를 수행하였다. Kang 등 (2015)은 사회네트워크분석과 테스트 마이닝기법을 이용하여 어떤 구단의 배구 경기력을 분석하였고, Bae 등 (2013)은 텍스트 마이닝을 이용한 기후변화 관련 식품분야 논문 초록에서 용어들의 출현빈도를 분석하였다. An과 Cho (2010)는 뉴스 기사 데이터를 이용하여 뉴스 내용이 주가 상승 또는 하락에 영향을 미치는지를 예측하는 알고리즘을 제안하였다. Jun과 Im (2014)은 소셜네트워크 분석을 활용한 생보사와 손보사의 대면/비대면 채널의 적합성 비교에 대해서 연구하였고, Lee와 Kim (2014)은 데이터마이닝 모형을 활용하여 호흡기질환의 주 요인을 선별하였다. Park와 Lee (2009)는 복합적 텍스트 분석을 이용한 포털 댓글에 관한 연구를 수행하였고, Oh와 Jin (2012)은 텍스트 마이닝을 이용한 쇼핑몰 구매후기 분석에 관한 연구를 수행하였다. Oh 등 (2010)은 텍스트 마이닝 분석을 통하여 어떤 패션회사의 고객 불만을 해소할 수 있는 방안에 대해서 연구하였고, Baek 등 (2015)은 텍스트 마이닝 분석 방법 중 Stylometry 방법에 대한 고찰과 이를 이용한 영화 흥행 예측에 관한 연구를 수행하였다. Jung (2010)은 텍스트 마이닝과 네트워크분석을 활용한 미래예측 방법 연구라는 주제로 정성적인 방법에 의해 작성된 제 3회 과학기술예측조사 (대조군)의 기술들과 정량적인 방법 (텍스트 마이닝)에 의한 과학기술 미래비전 (실험군)의 기술들을 비교 분석하였다. 이와 같이 다양한 분야에서 얻어진 방대한 비정형 데이터는 다방면으로 적용이 가능하다는 것을 알 수 있다.

영화시장에서는 일반적으로 사람들이 영화를 선택할 때, 네티즌들의 리뷰가 영화를 선택하는데 많은 영향을 미치고 있다. 즉, 네티즌들은 영화를 보기 전 관객들의 반응을 미리 확인하여 조금 더 만족할 만한 영화를 선택하여 관람하기를 원한다. 본 연구의 목적은 영화를 보고자 하는 네티즌들을 만족시키기 위해 영화의 리뷰와 몇 개의 변수를 이용하여 영화 흥행 여부를 판단할 수 있는 예측 모형을 개발하는데 있다.

본 연구의 구성으로 2절에서는 텍스트 마이닝에 대한 개괄적인 소개를 하고, 3절에서는 분석데이터에 대한 소개와 함께 연구방법에 대해서 소개한다. 4절에서 소개된 분석과정에서 얻어진 결과를 이용하여 다양한 분석을 수행한 후 분석 결과를 제시한다. 마지막 절에서는 분석결과와 연구의 한계점 및 향후 연구에 대해서 기술한다.

2. 분석 기법

2.1. 텍스트 마이닝

텍스트 마이닝은 문서상의 의미 있는 패턴과 관계를 찾고 이를 추출하는 것으로써 적절한 알고리즘을 이용하여 비정형화된 데이터를 정형화 데이터로 변환시켜 분석하는 일련의 과정이다. 텍스트 마이닝과 데이터 마이닝의 차이점은 데이터마이닝은 정형화된 데이터를 이용하고 텍스트 마이닝은 정형화되지 않은 데이터를 사용하는 것이다. 이러한 텍스트 마이닝의 등장으로 비정형화된 데이터를 텍스트간의 암묵적인 정보를 추출할 수 있으며, 비정형화된 텍스트 데이터들을 구분하여 이들 간의 연관성을 찾아 데이터를 클러스터링, 구조적 데이터와 결합을 통해 모델 구축 등을 수행한다 (SAS Institute INC, 2010).

2.2. 텍스트 마이닝 과정

텍스트 마이닝은 (1) '데이터 수집과정', (2) '용어 추출과정', (3) '정보 추출과정', (4) '정보 분석과정'의 4단계의 절차를 거친다 (Kang 등, 2015).

첫째, '데이터 수집과정'은 텍스트 마이닝의 첫 번째 단계로서 비정형 대규모 텍스트 데이터를 수집하는 단계이다. 데이터 수집과정에서 웹상에 있는 많은 양의 텍스트들을 한꺼번에 불러들일 수 있는 단계까지 발전되어 있다.

둘째, '용어 추출과정'은 문장의 단어, 규칙 등의 연관성을 고려하는 연관성 분석에 의해 단어들을 추출하여 관심이 있는 후보 단어를 만들어 내는 과정이다. 이러한 용어에 대해서 다양한 통계적 방법을 통해 전체를 대표하는 용어들을 추출하는 기법이다. 문장에서 용어들을 추출하는 방법은 TF (term frequency), DF (document frequency) 등이 있다. TF는 특정단어가 하나의 문서 (document)내에 나타난 용어를 추출하는 것이고, DF는 전체 문서 (global document) 세트 중에서 특정단어를 포함하는 문서의 용어를 추출하는 것이다.

셋째, '정보 추출과정'은 문서를 검색하는 것이 아니라 문서 내에서 필요한 정보를 추출 과정이다. 영화와 관련하여 예를 들면, 영화명, 주연배우, 배급사 등과 같이 상세 정보를 포함하는 용어를 추출하는 과정이다.

넷째, '정보 분석과정'은 세 번째 과정에서 얻어진 최종 키워드들에 대해서 '빈도', '분류', '클러스터링', '컨셉링크' 기법 등을 이용하여 유용한 정보를 도출해 내는 과정이다. '빈도'는 가장 많이 얻어지는 키워드가 무엇인지에 대한 정보를 도출해 내고, '분류'는 앞서 추출된 텍스트의 내용에 따라 문서들을 범주화 시켜주는 과정이다. 즉, 주어진 신문 텍스트 문서가 스포츠 분야인지, 정치 분야인지 등을 텍스트에 따라 분류하는 것을 의미한다. '클러스터링'은 문서에 포함되어 있는 추출된 단어들을 유사도에 따라 여러 개의 텍스트 집단으로 군집화 시켜주는 과정이다. '컨셉링크'는 어떤 특정한 키워드를 중심으로 또 다른 키워드들 간의 관계를 파악하는 기법이다.

2.3. 오피니언 마이닝

오피니언 마이닝 (opinion mining)이란 소셜미디어 또는 웹사이트에 나타난 의견을 분석해 유용한 정보로 만드는 기술로 네티즌들의 의견을 객관적인 정보로 바꿔주는 것이다. 예를 들면, 제품의 상품평, 영화의 리뷰 등과 같은 정보 중에서 유용한 정보를 찾고, 네티즌들의 생각과 표현을 이용하여 어떤 규칙성을 찾아내어 새로운 정보를 발굴하고 탐사하는 방법이다 (Yu 등, 2013). 텍스트 마이닝에 속하는 오피니언 마이닝은 주로 다양한 소셜 미디어 콘텐츠로부터 상품 및 서비스의 선호도, 사회적 사건이나 정치 이슈 등에 대한 대중들의 의견을 분석하는데 적용되고 있다. 특히 언어적 자원을 구축하는 연구에서는 개별 언어의 내용을 대상으로 어휘 또는 어의 수준에서 긍정, 중립, 부정 등의 평가를 정리해 놓은 감성사전을 사용하며, Wordnet의 각 어휘에 긍정, 중립, 부정 값을 태깅 (tagging)한 Senti Wordnet 기반 연구가 있다 (Yu 등, 2013). Yune 등 (2010)은 오피니언 마이닝 기술을 이용한 효율적인 상품평을 검색하는 기법을 연구하였다. 그들은 네티즌들의 상품평을 의도에 따라 순위를 결정하는 기법을 제안하였는데 네티즌들이 검색한 검색어뿐만 아니라 상품평 내에 주관적인 의견의 포함 여부 및 감정 극성의 엔트로피 등을 고려하여 상품평의 가치를 판단하였다. 본 연구에서도 영화와 관련된 네티즌들의 리뷰 데이터에서 감성단어를 추출 또는 긍정과 부정으로 표현된 주요 리뷰를 추출하는 등 오피니언 마이닝 개념을 사용하였다.

3. 연구 방법

3.1. 분석데이터

본 연구에서 사용된 데이터는 포털 사이트 'D'사와 '영화진흥위원회'의 영화에 대한 55,028개의 리뷰 데이터, 평점평균 그리고 스크린 수이다. 세부적으로 포털사이트 'D'사에서 제공된 2013년부터 2014년까지 상영된 영화 47편의 평점평균과 댓글 55,028개, 영화진흥위원회에서 제공하고 있는 47편의 영화에 대한 스크린 수로 구성되어 있다. 분석데이터를 구성하기 위해서 먼저 원시데이터인 47편의 영화에 대한 텍스트 문서에서 모든 영화에서 공통적으로 나오는 용어들 중에서 빈도수가 100 이상인 용어를 추출하고 이들 중에서 상위 10개를 선택하여 원시데이터를 1차 가공하였다.

Table 3.1 Variable to obtain explanation variable

variable	explanation
score	film grading (0∼ 10point)
screen	screening number of a newly released film
SVD1	variable for 'interesting'
SVD2	variable for 'emotion'
SVD3	variable for 'uninteresting'
SVD4	variable for 'boring'
SVD5	variable for 'satisfaction'
SVD6	variable for 'attraction'
SVD7	variable for 'sympathy'
SVD8	variable for 'impressive'
SVD9	variable for 'disappointment'
SVD10	variable for 'recommendation'

통상적으로 텍스트 마이닝에서는 특이값 분해 (singular value decomposition; SVD)을 수행하여 분석하고자 하는 데이터 셋을 얻게 된다. 47×10 의 1차 가공된 데이터 셋들을 특이값 분해를 수행해서 얻어진 고유벡터들을 SVD1부터 SVD10이라고 하자. 여기서 SVD 값들은 선형 결합된 고유벡터들이다. 이제 본 연구의 목적인 영화 흥행 여부의 모형을 얻기 위해 로지스틱 회귀분석에 적용될 데이터 셋은 47개의 영화와 SVD1부터 SVD10까지의 변수와 평점평균과 스크린 수 2개의 변수를 추가하여 설명변수로 하고 영화 흥행여부를 종속변수로 구성된 크기 47×13 행렬로 최종 데이터 셋으로 구성하였다. 여기서 47편의 영화는 2013년부터 2014년까지 흥행영화 34편과 비 흥행영화 13편을 선정하였다. 영화 흥행 여부는 영화 누적관객 수가 30만 명을 기준으로 하였다. 영화 흥행 여부에 대한 기준은 투자 대비 수입 등의 문제로 영화 흥행 여부에 대한 절대적이고 객관적인 기준이 없기 때문에 본 연구에서는 영화 관람 관객 수가 현저히 떨어지는 30만 명을 기준으로 하였다. 변수들에 대한 설명은 Table 3.1에 주어져 있다.

3.2. 분석 내용 및 방법

통상적으로 텍스트 마이닝에서 회귀분석을 수행하는 과정에서 특이값 분해를 이용한다. 특이값 분해 결과를 통해 얻어진 고유벡터 (빈도로 구성된 변수)을 설명변수로 사용한다 (SAS Institute INC, 2010). 본 연구에서는 영화를 행으로 하고 단어를 변수로 상정하여 열로 하는 행렬을 만들고 이를 특이값 분해를 이용하여 얻어진 고유벡터를 설명변수로 하고, 영화의 흥행여부를 종속변수로 하여 영화 흥행 여부를 예측하는 모형을 구축한다. 이를 위하여 3.1절에 기술한 분석데이터를 가지고 로지스틱 회귀분석을 수행한다. 여기서 변수선택법으로 단계별변수선택법을 사용하였다. 분석 결과는 4.4절의 로지스틱 회귀분석 결과에서 소개한다.

본 연구에서는 SAS Enterprise Miner 13.1의 Text Miner tool을 사용하여 분석하였다.

3.3. 분석 과정

3.3.1. 데이터 가공 및 정제

텍스트 마이닝의 첫 번째 단계로서 분석하고자 하는 데이터를 수집하여 연구 목적에 맞도록 가공 및 정제하는 단계로서 매우 중요한 단계이다. 수집된 원시 데이터는 한국어 문법 (명사, 동사, 형용사 등)에 맞지 않게 분류가 되어 있기 때문에 분석 목적에 맞도록 가공 및 정제를 해야 된다. 예를 들면, '재밋다', '잼있따', '재미나다'라는 단어를 하나의 동일어로 '재미있다'라는 등의 가공하는 작업이 필요하다. 데이터 가공 및 정제 단계는 분석에 필요한 속성에 맞게 원시데이터를 가공하고 정제하는 단계이다. Figure 3.1은 본 연구에서 사용된 원시데이터를 텍스트 마이닝 분석을 위한 원시 데이터의 가공 및 정제하는 과정을 보여 주고 있다.

3.3.2. 용어 추출과정

텍스트 마이닝의 두 번째 단계로서 분석에 필요한 용어를 추출하는 과정이다. 이를 위해서 텍스트 필터 노드에서 분석할 문서와 단어의 총 개수를 줄이는 작업을 수행할 수 있다. 즉, 문장 내에서 어떤 특정한 단어가 10개 이상 포함하는 댓글만을 선정할 수 있다. 예를 들면, 약 5만 5천개의 댓글 중에서 용어 필터 옵션을 이용하여 최소 문서수를 10으로 지정하면 어떤 특정 단어의 수가 9개 이하를 포함하고 있는 문서는 제거된다. 본 연구에서는 전기한 과정을 거쳐 상위 단어의 수는 20,000개로 제한하였다. 추출된 단어와 빈도수의 결과는 Table 3.2에 주어져 있다. 3.1절의 분석데이터에서 기술된 변수들은 43개 영화에서 공통적으로 나오면서 빈도가 100개 이상인 용어들 중 상위 10개이다.

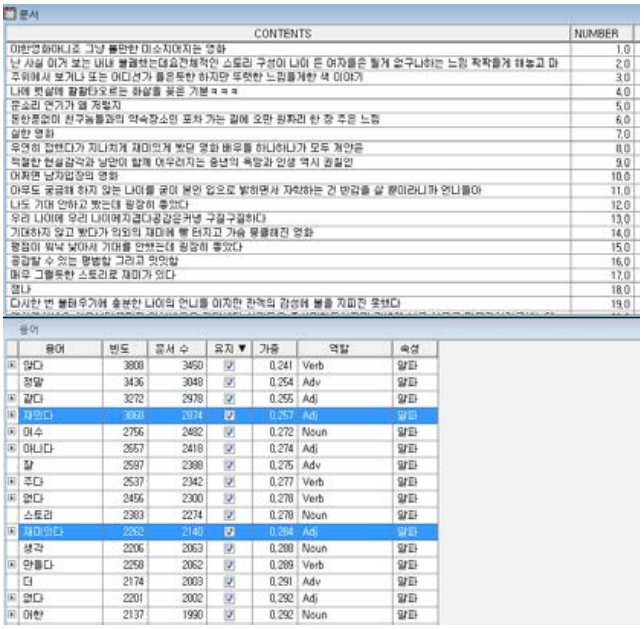

Figure 3.1 Refinement process in text miner

Table 3.2 Frequency of extracted words

term	frequency		term	frequency
interesting	7,253		transparent	519
emotion	2,878		pleasure	386
laugh	2,051		sad	382
boring	2,254		recommendation	528
uninteresting	1,288		stylish	398
⋮	⋮	⋮	⋮	⋮
acting	6,631		satisfy	417
grading	1,957		exorbitant	118
actor	3,892		commonness	122
action	1,052		masterpiece	106
drama	472		scale	111

3.3.3. 정보 추출과정

텍스트 토픽 뷰어 분석을 위해 텍스트 토픽 기능을 이용하여 텍스트 속에서 공통된 화제를 추출해 냄으로써 주제에 따라 관련 규정, 문서 항목설명 및 다양한 아이디어를 찾을 수 있다. 텍스트 토픽 뷰어 분석을 통해 다양한 정보를 추출할 수 있다. 텍스트 토픽 뷰어 분석에서 사용되는 옵션들은 다음과 같다. (1) 단일어 토픽 개수 (Number of single-term topics)는 하나의 단어로 여러 문서를 분류한다. 예를 들면, '액션'이라는 단어로 문서를 분류할 수 있다. (2) 다중어 토픽 개수 (Number of multi-term topics)는 여러 개의 단어를 지정하여 문서를 분류한다. 예를 들어 3을 지정하면 '감동', '재미있다', '괜찮다'의 세 단어로 문서를 분류할 수 있다. (3) 토픽 상관관계 (Correlated topics)는 토픽들 사이에 상관관계를 지정할 수 있다. 이에 대한 결과는 군집분석 결과에서 제시한다.

3.3.4. 정보 분석과정

텍스트 마이닝의 네 번째 단계로서 정보분석 단계는 3.3.2절의 용어 추출과정과 3.3.3절의 정보 추출과정에서 얻어진 각 종 정보를 활용하여 '워드 클라우드', '컨셉링크 (Concept Link)', '분류 및 군집분석' 등을 통하여 유용한 정보를 도출해 내는 과정이다. 이 과정에서 얻어진 대한 결과들은 4절의 분석결과에 주어져 있다.

4. 분석 결과

4.1. 워드 클라우드

분석과정에서 추출된 단어들의 빈도를 보면, 재미있다 (7,327), 재미없다 (668), 후회한다 (276) 등 영화평에 대한 감정동사와 명사들이 많이 추출되었다. 즉, 이러한 감정동사와 명사들이 영화 흥행여부에 대한 정보를 많이 가지고 있다고 할 수 있다. Figure 4.1은 워드 클라우드 (Word Cloud) 결과이다. 워드 클라우드는 빈도가 높고 핵심어일수록 큰 글씨로 중심부에 표현되며, 어떤 말을 하고 있는지 한 눈에 볼 수 있는 유용한 비주얼 분석도구이다. 본 연구에서 얻어진 워드 클라우드 결과에서와 같이 '스토리', '재미있다', '평점', '아쉽다' 등과 같은 단어들이 주요한 정보로 얻어졌다.

Figure 4.1 Word Cloud

4.2. 컨셉링크

단어들 간의 관계를 시각적으로 보여주는 것이 컨셉링크이다. 연결 분석을 통해 다양한 결과를 보여준다. 문서 내에 단어들 간의 비율을 통해 특정단어를 기준으로 해서 선의 굵기에 따라서 중요성을 나타낸다. Figure 4.2에서 '재미있다'는 연기, 스토리, 구성과 연관성이 높은 것으로 나타났다. 이것은 영화에서 재미라는 것이 스토리, 구성, 연기, 배우 등 다양한 요소에서 보인다고 할 수 있다.

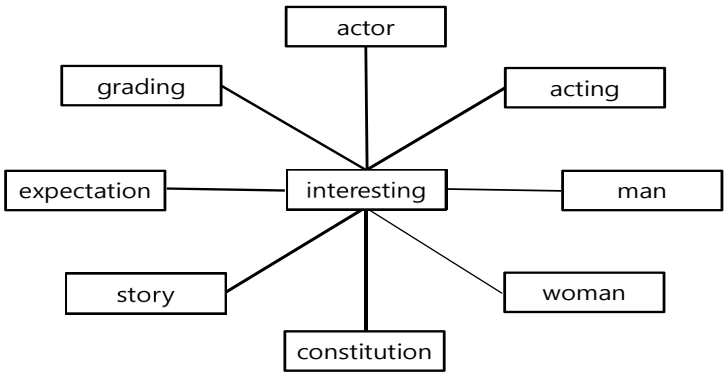

Figure 4.2 Concept Link

4.3. 군집분석

먼저 그룹 1에서는 '수상한 그녀', '신이 보낸 사람', '플랜맨', '명량', '피끊는 청춘'이라는 영화가 포함되어 있다. 이 영화들의 특징들은 장르에서 코미디라 '재미있다', '괜찮다'라는 주요단어들이 묶였다. 코미디는 아니지만 2014년 1,700만 명이라는 최고의 흥행작의 '명량'이 그룹 1에 포함 되어 있다.

그룹 2에서는 '인간중독', '가시', '좋은 친구들', '남자가 사랑할 때' 포함되었다. 사랑스러운 스토리로 주로 멜로, 로맨스 장르의 영화가 그룹이 되었다. 주요단어들은 '사랑', '결말', '아쉽다'이다.

그룹 3에서는 '방황하는 칼날', '또 하나의 약속', '관능의 법칙', '표적', '우는 남자'로 구성되었다. '눈물', '스토리', '울다'라는 단어가 묶였는데 영화들의 특징을 보면 실화를 바탕으로 가슴 아픈 스토리

로 구성되어 있다.

그룹 4에서는 '살인자', '끝까지 간다', '터널 3D', '소녀괴담', '경주' 포함되었다. 그룹 4는 공포, 스릴러의 장르로서 '무섭다', '스릴러', '연기력'의 단어들로 구성되어 있다.

그룹 5에서는 '하이힐', '찌라시', '군도', '역린'의 영화로 액션 장면이 많은 영화로 구성되어 있다. 그룹 5와 관련되어 있는 단어들은 '감독', '연출', '장면'의 단어로 구성이 되어있다.

그룹 6에서는 대체적으로 평점이 낮은 영화들이 그룹이 되었다. 영화는 '신의 한수', '몬스터', '우는 남자', '황제를 위하여', '표적'으로 구성되어 있고, 단어는 '액션', '평점', '아쉽다'로 구성되어 있다. 이를 표로 정리한 결과가 Table 4.1에 주어져 있다.

군집의 특징을 영화 장르로 보면 군집 1은 코미디, 군집 2는 멜로 및 로맨스, 그룹 3은 드라마, 그룹 4는 공포, 그룹 5는 액션, 그룹 6은 액션 및 드라마로 구분할 수 있었다.

Table 4.1 Result of clustering analysis

group	film name	key word (topic)
1	Miss granny, The apostle: He was anointed by God, The plan man, Myeongnyang, Hot young bloods	interesting, passable
2	Obsessed, Innocent thing, Good friends, Man in love	love, ending, miss
3	Broken, Another family, Law of sensory, The target, No tears for the dead	tear, story, cry
4	Murderer, A hard day, Tunnel 3D, Mourning grave, Gyeongiu	terrible, thriller, acting
5	Man on high heels, Tabloid truth, Kundo, The fatal encounter	direction, production, scene
6	The divine move, The monster, No tears for the dead, For the emperor, The target	action, grading, miss

4.4. 로지스틱 회귀분석

3절에서 기술한 분석데이터를 이용하여 영화 흥행 예측을 위해 로지스틱 회귀분석을 실시한 결과 Table 4.2에 주어져 있다.

Table 4.2 Result of logistic regression analysis (proposal method)

Parameter	target	DF	Estimate	SE	wald chi-square	pr>chisq
Intercept	1	1	-4.009	1.553	6.66	0.010
SVD1	1	1	0.117	0.050	5.56	0.021
SVD3	1	1	0.211	0.103	4.21	0.040
score	1	1	0.079	0.032	6.21	0.013

Table 4.2를 보면 단계적 변수선택법에 의해서 SVD1, SVD3 그리고 평점평균 변수들이 선택되어 졌고, 모든 변수들이 유의수준 0.05하에서 유의한 변수로 얻어졌다. 본 연구에서 제안한 모형에 의한 예측력의 결과는 Table 4.3에 주어져 있다. Table 4.3에서 실제로 흥행된 영화 34개 중 흥행의 결과로 예측된 결과가 32개, 비 흥행된 영화 13개 중에서 비 흥행으로 예측된 결과가 13개로 얻어져 정분류율은 95.74%로 얻어졌다. 여기서 절단값 (cutofff point)은 0.5로 얻어졌으며, 이 값은 분류 기준값별 정분류 행렬을 근거로 결정하였다.

Table 4.3 Prediction result

comparison		prediction		total
		performance	nonperformance	
real	performance	32	2	34
	nonperformance	0	13	13
	total	32	15	47

Table 4.2의 예측결과, 흥행된 영화가 비흥행으로 분류된 두 편의 영화는 '신이 보낸 사람 (누적관객 수: 약 42만 명)'와 '하이힐 (누적관객 수: 약 34만 명)'이었다. 잘못 분류된 이유는 누적관객 수가 다른 흥행 영화들보다 다소 적었기 때문인 것으로 판단된다.

본 연구에서 제안한 모형을 이용하여 미개봉작 영화인 '기술자들'에 대한 흥행 여부를 예측 (2014년 12월 18일 기준)한 결과 '흥행'으로 얻어졌다. 실제 개봉일이 2014년 12월 24일이고, '기술자들'의 영화는 관객 수가 256만 명을 넘어 실제로 흥행을 하였다.

5. 결론 및 향후 연구과제

본 연구에서 사용한 데이터는 포털 사이트 'D'사의 2013에서 2014년까지 영화 47편에 대한 55,028개의 리뷰와 영화 평점에 관한 데이터, 그리고 '영화진흥위원회'의 스크린 수에 대한 데이터로 구성되어 있다. 분석 방법으로는 SAS Enterprise Miner 13.1의 Text Miner tool을 사용하였다.

워드 클라우드 결과, '스토리', '재미있다', '평점', '아쉽다' 등과 같은 단어들이 주요 단어로 얻어졌다. 컨셉링크에 대한 결과는 '재미있다'의 단어는 연기, 스토리, 구성과 연관성이 높은 것으로 나타났다. 군집분석 결과는 총 6개 그룹으로 군집으로 얻어졌다. 군집의 특징을 영화 장르로 보면 군집 1은 코미디, 군집 2는 멜로 및 로맨스, 그룹 3은 드라마, 그룹 4는 공포, 그룹 5는 액션, 그룹 6은 액션 및 드라마로 구분할 수 있었다.

본 연구에서 사용된 설명변수를 도출해 내기 위한 데이터 구조는 47편의 영화에 대해서 10개의 변수로 47×10 의 행렬구조로 하여 1차 정제된 데이터를 구성하였다. 여기서 47편의 영화는 2013년부터 2014년까지 흥행영화 34편과 비 흥행영화 13편을 선정하였다. 그리고 본 연구에서의 목적을 달성하기 위해서 1차 정제된 47×10 의 데이터 셋을 이용하여 특이값 분해를 적용하여 고유벡터로 구성된 10개의 SVD 변수를 만들었다. 이 변수와 2개의 변수 (평점평균과 스크린 수)를 추가하여 설명변수로 하고 영화 흥행여부를 종속변수로 하여 최종 분석데이터 셋을 구성하였다. 단계적 변수선택법을 이용한 로지스틱 회귀분석 결과 얻어진 변수는 SVD1, SVD3 그리고 평점평균 변수가 얻어졌고, 이 변수들은 모두 유의수준 0.05하에서 유의하였으며 본 연구에서 제안한 영화 흥행 예측모형을 얻었다. 본 연구에서 제안한 예측 모형의 정분류율은 95.74%로 얻어졌다.

본 연구의 결과는 표본을 어디에서 얼마만큼 텍스트 데이터를 수집하느냐에 따라 분석결과가 달라질 수 있다는 단점이 있다. 그리고 본 연구에서 텍스트 마이닝을 위해 특이값 분해의 적용에 초점을 두었으나 보다 나은 영화 흥행 예측 모형을 도출해 내기 위해서는 다양한 독립변인 ('별점', '영화의 장르', '개봉시기' 등)을 고려해야 할 것으로 판단된다. 또한 본 연구에서는 2013년부터 2014년까지의 47편에 대한 영화리뷰 데이터를 이용한 연구결과이기 때문에 영화 흥행 여부에 대한 모형을 구축하는데 한계가 있다. 즉, 영화의 편수를 보다 많이 고려하는 것이 필요로 한다.

References

An, S. W. and Cho, S. B. (2010). Stock prediction using news text mining and time series analysis. *Journal of Computing Science and Engineering*, **37**, 77-82.

Bae, K. Y., Park, J. H., Kim, J. S. and Lee, Y. S. (2013). Analysis of the abstracts of research articles in food related to climate change using a text-mining algorithm. *Journal of the Korean Data & Information Science Society*, **24**, 1429-1437.

Baek, G. I., Kim, K. K., Choi, S. B. and Kang, C. W. (2015). Prediction for the Films Success using Stylometry. *Journal of the Korean Data Analysis Society*, **17**, 719-728.

Chun, H. J. and Leem, B. H. (2014). Face/non-face channel fit comparison of life insurance company and non-life insurance company using social network analysis. *Journal of the Korean Data & Information Science Society*, **25**, 1207-1219.

Jung, K. H. (2010). A study of foresight method based on text mining and complexity network analysis. *Korea Institute of S&T Evaluation and Planning*, Seoul.

Kang, B. U., Huh, M. K. and Choi, S. B. (2015). Performance analysis of volleyball games using the social network and text mining techniques. *Journal of the Korean Data & Information Science Society*, **26**, 1-12.

Kim, K. H. and Oh, S. Y. (2009). Methodology for applying text mining techniques to analyzing online customer reviews for market segmentation. *International Journal of Contents*, **9**, 272-284.

Lee. J. Y. and Kim, H. J. (2014). Identification of major risk factors association with respiratory diseases by data mining. *Journal of the Korean Data & Information Science Society*, **25**, 373-384.

Oh, S. W. and Jin, S. H. (2012). A study on analysis of internet shopping mall customers' reviews by text mining. *Journal of the Korean Data Analysis Society*, **14**, 125-137.

Oh, H. S., Cho, S. K., Kang, C. W. and Lim, D. S. (2010). Fashion Company's Claim Data Analysis Using Text Mining. *Journal of the Korean Data Analysis Society*, **12**, 297-306.

Park, H. W. and Lee, Y. O. (2009). A mixed text analysis of user comments on a portal site : The 'BBK Scandal' in the 2007 presidential election of south korea. *Journal of the Korean Data Analysis Society*, **11**, 731-744.

SAS Korea. (2010). G *etting Started with SAS Text Miner 4.2.*, SAS Siftware Korea Ltd.

Yu, E. J., Kim Y. S., Kim, N. K. and Jung, S. R. (2013). Predicting the direction of the stock index by using a domain-specific sentiment dictionary. *Journal of intelligence and information systems*, **19**, 95-110.

Yune, H. J., Kim, H. J. and Chang, J. Y. (2010). An efficient search method of product reviews using opinion mining techniques. *Journal of Computing Science and Engineering*, **16**, 222-226.

Journal of the Korean Data & Information Science Society
2015, **26**(6), 1259–1269

http://dx.doi.org/10.7465/jkdi.2015.26.6.1259
한국데이터정보과학회지

Study on prediction for a film success using text mining[† ‡]

Sanghun Lee[1] · Jangsik Cho[2] · Changwan Kang[3] · Seungbae Choi[4]

[1]Onsol Communication
[2]Department of Information Statistics, Kyungsung University
[34]Department of Data Information Science, Dongeui University

Received 21 July 2015, revised 4 September 2015, accepted 16 September 2015

Abstract

Recently, big data is positioning as a keyword in the academic circles. And usefulness of big data is carried into government, a local public body and enterprise as well as academic circles. Also they are endeavoring to obtain useful information in big data. This research mainly deals with analyses of box office success or failure of films using text mining. For data, it used a portal site 'D' and film review data, grade point average and the number of screens gained from the Korean Film Commission. The purpose of this paper is to propose a model to predict whether a film is success or not using these data. As a result of analysis, the correct classification rate by the prediction model method proposed in this paper is obtained 95.74%.

Keywords: Correct classification rate, opinion mining, singular value decomposition, text mining.

† This work was supported by Dong-eui University Grant (2015AA098).
‡ This work was reconstructed from a master's degree thesis of Sanghun Lee which is coauthor.
[1] Employee, Onsol communication, Seoul 153-803.
[2] Professor, Department of Informational Statistics, Kyungsung University, Busan 608-736, Korea.
[3] Professor, Department of Data Information Science, Dongeui University, Busan 614-714, Korea.
[4] Corresponding author: Professor, Department of Data Information Science, Dongeui University, Busan 614-714, Korea. E-mail: csb4851@deu.ac.kr